基于 GAN+XGBoost+LR 的个性化推荐方法

达钰鹏[1,2]　陈艳春[1]
[1]（石家庄铁道大学经济管理学院 石家庄 050043）
[2]（河北省人力资源和社会保障厅信息中心 石家庄 050071）

摘要：

[目的]解决长尾商品的推荐中存在的样本数据偏少，现有协同过滤法计算量大，性能难以满足需求的问题

[方法]出了基于 GAN+XGBoost+LR 的解决方案，通过协同过滤寻找初始样本数据，利用 GAN 生成更多样本数据训练 XGBoost+LR 模型，并针对不同模型的特点指定针对性的训练策略。

[结果]该方案在兼顾性能和精确度要求下，可以提高推荐模型的鲁棒性。

[局限]XGBoost 模型承担自动化特征工程能力有限。

[结论]基于 GAN+XGBoost+LR 的个性化推荐方法可以提高长尾商品的推荐的有鲁棒性。

关键词：机器学习；个性化推荐；对抗生成网络；XGBoost；长尾商品

分类号：TP393

Personalized Recommendation Method Based on GAN+XGBoost+LR

Da Yupeng[1,2], Chen Yanchun[1]
[1]（School of Economics and Management, Shijiazhuang Tiedao University, Shijiazhuang, 050043）
[2]（Information Center, Hebei Provincial Department of Human Resources and Social Security, Shijiazhuang, 050071）

Abstract：

[Objective] Solve the problem that there are few sample data in the recommendation of long tail merchandise, the existing collaborative filtering method has a large amount of calculation, and the performance can not meet the needs.

[Methods] A solution based on GAN+XGBoost+LR is presented. Initial sample data is searched through collaborative filtering, more sample data is generated by GAN to train XGBoost+LR model, and specific training strategies are specified according to the characteristics of different models.

[Results] This scheme can improve the robustness of the recommended model with both performance and accuracy requirements.

[Limitations]The XGBoost model has limited ability to undertake automated feature engineering.

[Conclusions]Individual recommendation method based on GAN+XGBoost+LR can improve the robustness of long tail recommendation.

Keywords： Machine learning; Personalized recommendation; Counter-generating network; XGBoost; Long Tail Goods

1 绪论

电子商务商品数量的无限性，使得用户搜索自己想要的商品变得愈加困难，

信息过载现象使得用户不得不花费巨大的精力从海量信息中进行选择[1]。推荐系统是解决信息过载的常见解决方案，通过学习用户历史行为偏好主动预测用户对未评分或反馈的物品潜在的反应，通过推荐列表等形式自动化、个性化地推荐给不同用户，从而帮助他们更快速地找到需要或喜欢的物品。因此，如何在有限的资源和较高的效率条件下完成对长尾商品的冷启动是一个热点问题[2]。

在个性化推荐方法发展过程中，商品的冷启动一直是一个重要的问题。可分为基于机器学习得方法和基于深度学习得方法。基于机器学习的方法又可以分为基于相似性的和基于模型的方法：基于相似性的方法是根据用户或物品之间的相似性进行预测，例如同一个作者或者主题的书籍，或者同一地区的男性用户等；基于模型的方法则是从用户对物品的点击或者购买行为数据构建预测模型。基于深度学习的方法则是利用深度学习的强大特征提取能力和非线性拟合能力来做推荐[3]。冷启动问题包括两个方面：新物品的冷启动和新用户的冷启动。对于新用户，他们没任何或者几乎没有对物品的历史评分信息、评论信息等等历史行为数据，以至于对于一些利用用户的历史行为来构建的推荐方法，它们是无法为新用户推荐有效的信息（即：用户冷启动）；对于一个新的物品，由于没有评论、浏览、交易、点击、评分等行为数据，这样就很难将这些物品推荐给对它们感兴趣的用户，从而导致物品的冷启动问题。

针对此问题，国内外很多学者提出了解决思路。例如，Zahid[4]等人提出基于邻域模型中使用的物品细节的归一化技术，以对冷启动问题的用户参与或用户喜好进行建模，该模型利用归一化技术，将新物品和新用户按其相关标签信息进行排序，并将这些信息映射到一个相同的隐含特征因子空间，以实现在用户和物品的基线评分上进行归一化，以帮助在少量数据的情况下解决冷启动问题；Tahmasebi[5]等人提出了一种基于剖面展开技术的混合推荐方法，缓解个性化推荐系统中存在的新用户冷启动问题，该模型首先提出一种机制通过添加一些额外的评分以创建一个比原来更密集的用户道具评分矩阵，并为目标用户提供进一步评分的评分配置文件以缓解推荐系统中的冷启动问题；Paled[6]等人提出了一种利用基于社区检测的交替最小二乘分解方法，该方法利用了 Louvain 算法和交替最小二乘算法的优点，并采用 Louvain 算法分析用户之间的关系，再采用交替最小二乘算法预测推荐来解决冷启动问题.Yadav[7]等人提出一种基于链接开放数据和社交网络特征的推荐系统方法，该方法通过构建基于链接开放数据（Linked Open Data, LOD）协作特征和基于社交网络的特征的用户档案来解决纯新用户冷启动问题。

总体上来说，其解决的思路分为两类，一类是挖掘商品或用户之间的关联性，将新商品同现有商品，新用户和现有用户关联起来，利用现有的推荐模型进行推荐；另一类是通过用户的社交网络信息、商品的标签信息等第三方数据，丰富新商品或新用户的数据，进而构建预测模型。本文针对长尾商品的冷启动问题，提出基于 GAN+XGBoost+LR 的个性化推荐方法。

2 基于 GAN+XGBoost+LR 的推荐方法设计

2.1 设计思路

针对长尾商品的个性化推荐，常见的方法是使用协同过滤，首先对商品进行分析，通过打标签、分类等方法丰富商品信息，而后进行商品的协同过滤找出相似的商品，与历史的购买用户进行关联，而后进行推荐。

协同过滤掉方法最大的好处是理论简单，易于实现，但面对长尾商品推荐时存在着以下几个问题：

第一，资源需求高。长尾商品最大的特点就是数量多，协同过滤算法需要维护一个巨大商品相似度矩阵，对内存和算力的要求成几何增长。

第二，模型更新缓慢。由于每次模型更新需要对矩阵内所有商品的相似度进行计算，其计算量相当惊人，面对实时型要求日渐提高点今天，已不能满足要求。

第三，冷启动效果差。长尾商品的信息较少，打标签、分类等手段依赖人工，工作量大，在冷启动阶段协同过滤效果差强人意。

综上可知，长尾商品的个性化推荐的要求有三个：一是资源需求不要太高，二是能更新速度要快，三是能比较好的缓解冷启动问题。基于业界成功的 XGBoost+LR 模型，本文提出利用 GAN+XGBoost+LR 的解决方案。

训练模型首先要解决的就是数据问题。这里的解决方法是，启动前理由协同过滤找出相似度高的历史商品，以该数据为基础利用 GAN 生产更多样本数据进行分析；在商品启动后，基于用户的反馈数据利用 GAN 生产更多样本数据，进行模型再训练。

在模型选择和训练策略上。这里采用了 XGBoost+LR 的模型，利用 XGBoost 进行特征工程，利用 LR 完成最后的 CTR 值输出，两个模型采用不同的训练策略，XGBoost 模型使用 GAN 模型生成的数据进行不定期在线更新，LR 模型利用反馈真实数据增量更新，定期对 GAN+XGBoost+LR 进行离线更新。采取这样的策略是因为 LR 模型训练速度最快，XGBoost 次之，GAN 最慢，为了兼顾速度和精度需要采用不同的训练策略。具体算法流程如图 1 所示：

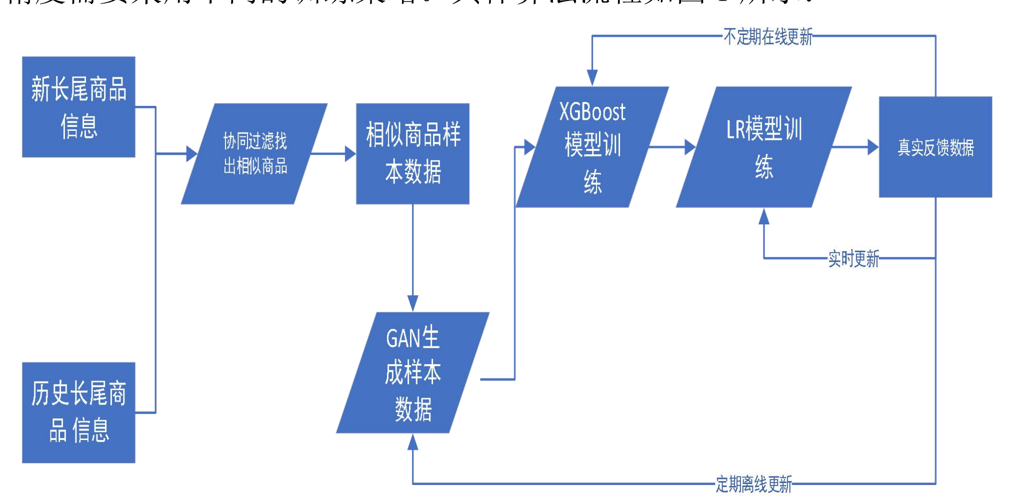


图 1 算法流程图

在整个流程中，需要计算资源较多的步骤是协同过滤和 GAM 生产样本数据，其中协同过滤仅使用一次，GAN 按照需要定期使用，避免带来过多的计算压力；而相对较慢的 XBGoost 模型则次用了不定期在线更新和定期离线更新结合的方式，不定期在线更新主要是针对模型参数的微调，在已有参数基础上进行上下浮动微调；LR 模型由于运算速度较快可实时使用真实数据进行训练；而 GAN+XGBoost+LR 整个模型店更新，则根据实际情况定期进行离线训练。

2.2 算法实现

具体算法实现基于 python 语言，使用 anaconda 软件包、scikt-learn 软件包和 ydata-synthetic 软件包，利用 anaconda 进行数据处理，scikt-learn 软件包实现 LR 模型，pyxgboost 包实现 xgboost 模型，ydata-synthetic 软件包实现 GAN 模型:

XGBoost+LR 模型实现步骤：
1.读取数据集数据为 df
2.按照最大最小值进行标准化并分割位训练集和测试集
3.用训练集数据数据训练 XGBoost 决策树模型
4.将训练模型店叶子节点特征导出
5.用叶子节点特征处理数据并训练 LR 模型
RNN 模型进行对比验证步骤：
定义 RNN 回归模型:

```
def  RNN_regression(units):
     opt  =  Adam(name='AdamOpt')
     loss=MeanAbsoluteError(name='MAE')
     model  =  Sequential()
     model.add(GRU(units=units,
     name=f'RNN_1'))
     model.add(Dense(units=21,
activation='sigmoid',name='OUT'))
      model.compile(optimizer=opt,  loss=loss)
      return  model
ts_real  =  RNN_regression(12)
early_stopping  =  EarlyStopping(monitor='val_loss')
real_train=ts_real.fit(x=X_stock_train,y=y_stock_train,validation_dat
a=(X_stock_test,y_stock_test),epochs=200,batch_size=128,callbacks=[ea
rly_stopping])
ts_synth  =  RNN_regression(12)
synth_train  =  ts_synth.fit(x=X_synth_train,
y=y_synth_train,
validation_data=(X_stock_test,  y_stock_test),
epochs=200,
batch_size=128,
callbacks=[early_stopping])
```


2.3 实验环境与数据

使用设备为个人 PC 计算机，处理器为 AMD Ryzen 5 2600X　3.60 GHz，内存 32.0 GB，250GB SSD 固态影片，HP GTX750Ti 4GB 显卡。数据采用阿里云天池平台 IJCAI-18 阿里妈妈搜索广告转化预测中第一赛季的 A 榜训练数据[8]，文件名为 round1_ijcai_18_train_20180301.txt，文件大小 529 MB，共有 478087 条数据，为 2018 年 9 月 17 日至 2018 年 9 月 24 日共 8 天的数据。数据集分为 5 个表，每个表各个字段的解释如下表 1 至表 5。

表 1：基础数据表

字段	解释
instance_id	样本编号，Long
is_trade	是否交易的标记位，Int 类型；取值是 0 或者 1，其中 1 表

	示这条样本最终产生交易，0 表示没有交易
item_id	广告商品编号，Long 类型
user_id	用户的编号，Long 类型
context_id	上下文信息的编号，Long 类型
shop_id	店铺的编号，Long 类型

表 2：广告商品信息表

字段	解释
item_id	广告商品编号，Long 类型
item_category_list	广告商品的的类目列表，String 类型
item_property_list	广告商品的属性列表，String 类型
item_brand_id	广告商品的品牌编号，Long 类型
item_city_id	广告商品的城市编号，Long 类型
item_price_level	广告商品的价格等级，Int 类型；取值从 0 开始，数值越大表示价格越高
item_sales_level	广告商品的销量等级，Int 类型；取值从 0 开始，数值越大表示销量越大
item_collected_level	广告商品被收藏次数的等级，Int 类型；取值从 0 开始，数值越大表示被收藏次数越大
item_pv_level	广告商品被展示次数的等级，Int 类型；取值从 0 开始，数值越大表示被展示次数越大

表 3：广告商品信息表

字段	解释
user_id	用户的编号，Long 类型
user_gender_id	用户的预测性别编号，Int 类型；0 表示女性用户，1 表示男性用户，2 表示家庭用户
user_age_level	用户的预测年龄等级，Int 类型；数值越大表示年龄越大
user_occupation_id	用户的预测职业编号，Int 类型
user_star_level	用户的星级编号，Int 类型；数值越大表示用户的星级越高

表 4：上下文信息表

字段	解释
context_id	上下文信息的编号，Long 类型
context_timestamp	广告商品的展示时间，Long 类型；取值是以秒为单位的 Unix 时间戳，以 1 天为单位对时间戳进行了偏移
context_page_id	广告商品的展示页面编号，Int 类型；取值从 1 开始，依次增加；在一次搜索的展示结果中第一屏的编号为 1，第二屏的编号为 2
predict_category_property	根据查询词预测的类目属性列表，String 类型；property_B 取值为-1，表示预测的第二个类目 category_B 没有对应的预测属性

表 5：店铺信息表

字段	解释
shop_id	店铺的编号，Long 类型
shop_review_num_level	店铺的评价数量等级，Int 类型；取值从 0 开始，数值越大表示评价数量越多
shop_review_positive_rate	店铺的好评率，Double 类型；取值在 0 到 1 之间，数值越大表示好评率越高
shop_star_level	店铺的星级编号，Int 类型；取值从 0 开始，数值越大表示店铺的星级越高
shop_score_service	店铺的服务态度评分，Double 类型；取值在 0 到 1 之间，数值越大表示评分越高
shop_score_delivery	店铺的物流服务评分，Double 类型；取值在 0 到 1 之间，数值越大表示评分越高
shop_score_description	店铺的描述相符评分，Double 类型；取值在 0 到 1 之间，数值越大表示评分越高

本质上来说，这个数据集是一个预测 CVR 的问题，进一步对 item_id 数据进行分析，计算每个 item_id 在数据中出现的次数，即被展示的次数，可以提现该商品的热度高低，按照次数由高到低排序后绘制折线图，如图 2 所示：

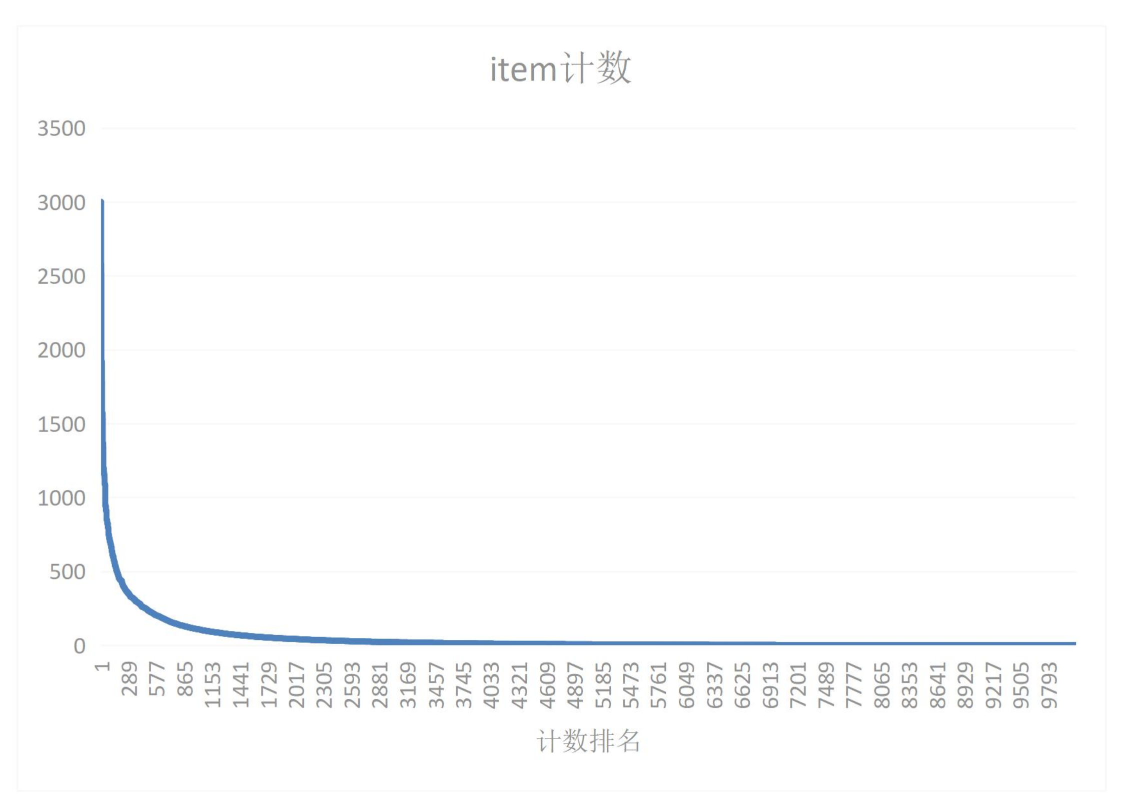


图 2 item_id 计数分布

通过以上图标可知大量的 item_id 出现的次数很低，符合长尾现象，我们由高到低的计算每个 item_id 出现次数占总次数多百分比，并同时计算累计百分比，绘制折线，如图 3 所示：

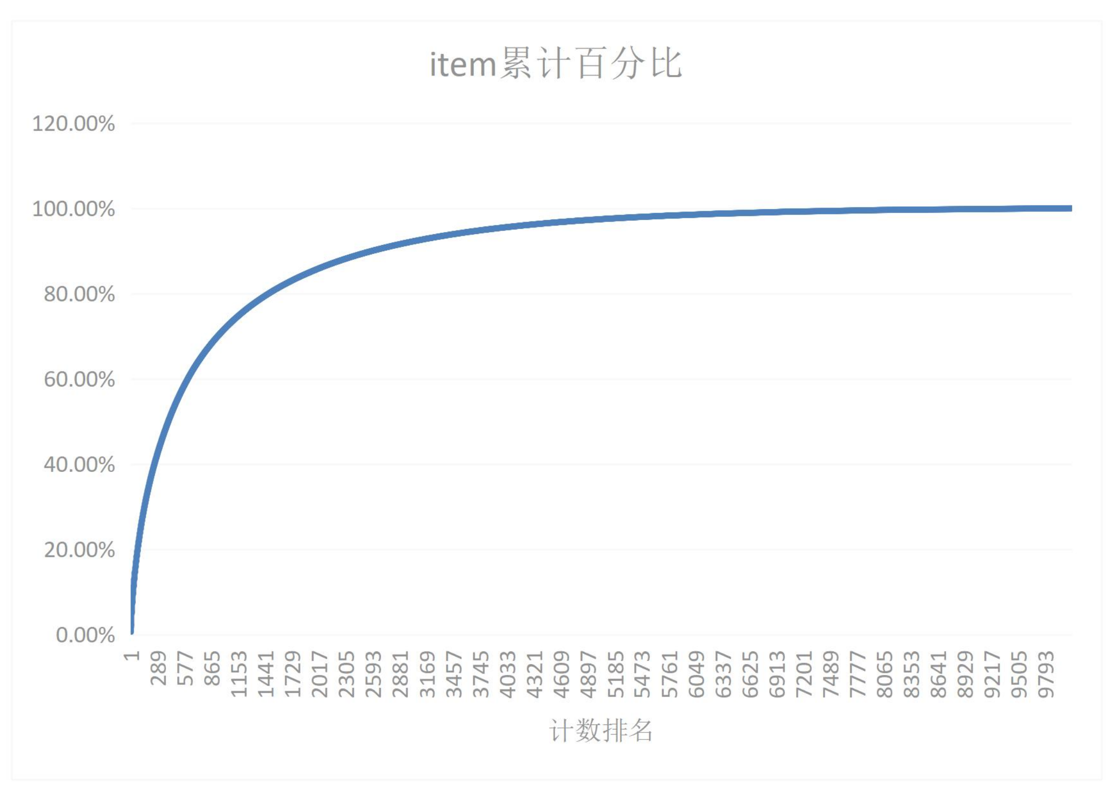


图 3 item_id 累计百分百

如上两图可知，该数据集中的商品分布符合长尾现象，进过进一步统计可知，

排名前 1460 位的商品展示次数共有 381311 次，即 14.49%的商品展示次数占了总数 79.75%，商品冷热不均情况显著，冷热数据的分界点在 66 次，即出现高于 66 次的为热数据，低于 66 次的为冷数据，我们提取这部分冷数据作为实验数据。

这部分数据共有 96827 条，涉及商品数量 8615 个，平均每个商品展示 11.24 次，最少展示 1 次，count 详细值统计如图 4 所示：

	count
count	8615.000000
mean	11.239350
std	14.540738
min	1.000000
25%	2.000000
50%	4.000000
75%	15.000000
max	66.000000

图 4 count 值统计情况

2.4 实验数据分析和处理

通过前面的统计分析可知，25%的商品展示次数为 1-2 次，50%的商品展示次数为 1-4 次，75%的商品展示次数为 1-15 次，25%的商品展示次数高于 15 次，发生交易的样本数为 1595 个，占总数的 1.675%，未发生交易的样本数为 95232 个，占总数的 98.353%。由于本文的重点在于 GAN 生成的数据能否提升 XGBoost+LR 在 CVR 预计的效果，故不在特征工程上花太多功夫，稍作处理后使用 XGBoost+LR 进行模型训练，logloss 值是 0.072366，下面我们尝试使用基于 keras 的 GAN 生成新的样本数据集，实现对现有数据集规模的扩充。

该数据集数据可以看作一个时间序列，这里采用 TimeGAN 模型进行训练，TimeGAN 认为时序数据应该有两种特征，第一个是静态特征，不会因为时间而改变的特征，例如本文所用数据中的商品价格、用户性别等数据；第二个是时态特征，该特征随时间而改变，例如生商品的广告展示次数、收藏等数据[9]。由于本文使用数据正负样本存在明显的不平衡性，为了提高准确性，用正负样本分别训练 TimeGAN 模型生成数据和真实数据进行对比，并通过 PCA（主成分分析）和 t-SNE（t-分布式随机邻居嵌入）可视化数据分布，然后分别用生成数据和真实数据训练 RNN 模型，比较生成数据和真实数据训练模型的差距。

（1）正样本生成数据。图 5 为正样本的 PCA 和 t-SNE 可视化散点对比图，灰色点为真实数据，红色点为生成的数据，图 6 为生成数据与真实数据折现对比图。

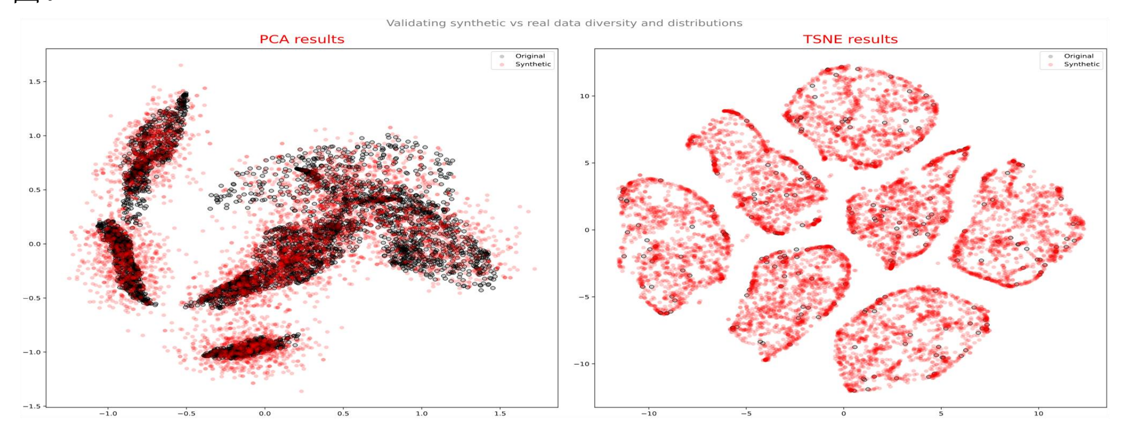


图 5 正样本 VS 正样本生成数据可视化对比图

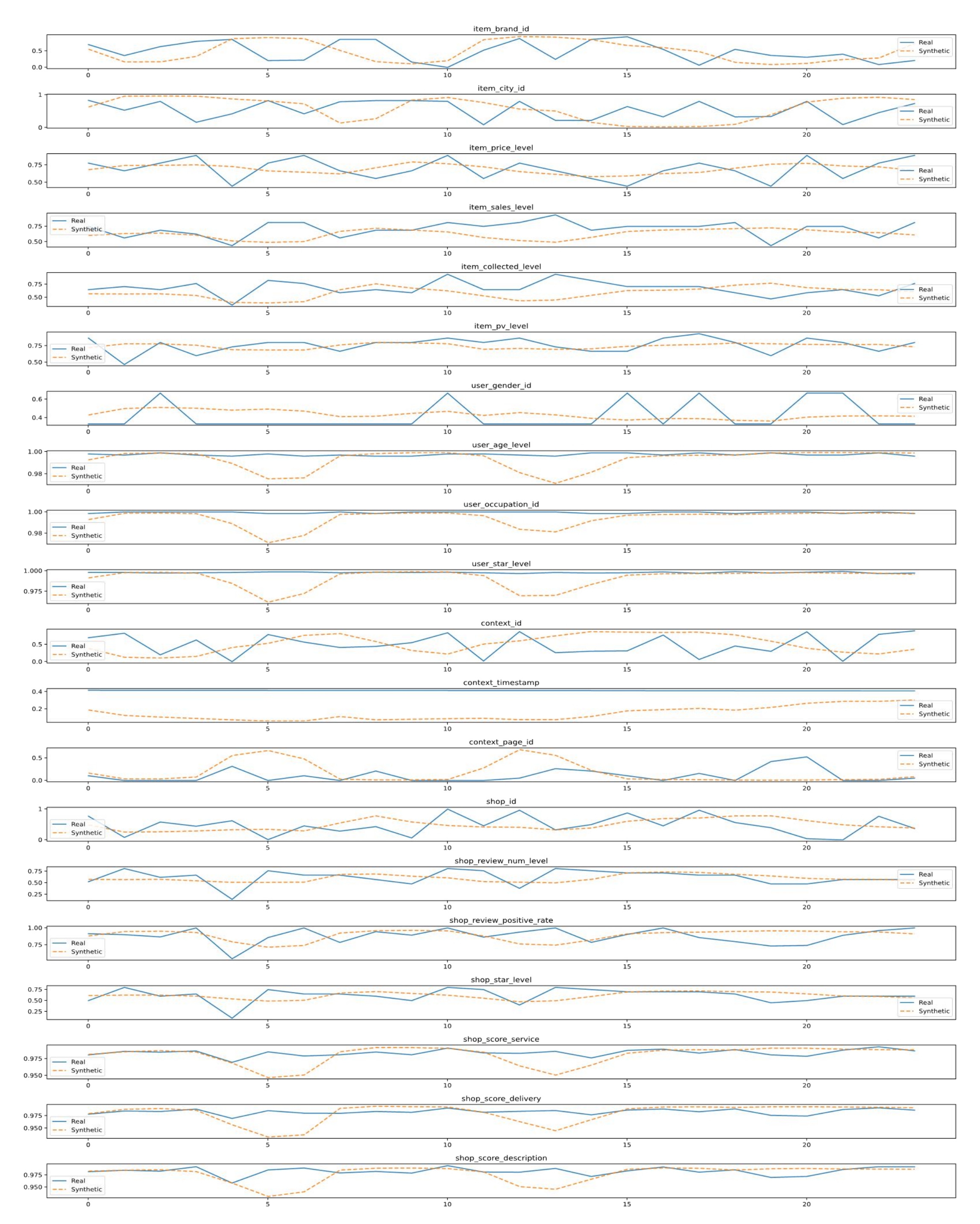


图 6 正样本 VS 正样本生成数据对比折线图

为了验证生成数据和真实数据在模型上的差异，利用 RNN 模型进行对比验证。比较指标主要有 MSE（均方误差）、MAE（平均绝对误差）和 MSLE（均方对数误差），为了保证数据可靠性，使用同一份生成数据反复训练 RNN 模型十轮，求均值，结果如表 6 所示：

表 6：正样本 VS 正样本生成数据模型表现

十轮均值	MSE	MAE	MSLE
真实数据	0.0315846	0.1090066	0.0150415
生成数据	0.0333358	0.118209	0.0157859
性能差距	-5.54%	-8.44%	-4.95%

由此表可知，正样本生成数据在实际模型中和直接用真实数据的差距较小，性能损失在 4.95%-8.44%之间。

（2）负样本生成数据。负样本生成数据采用和正样本一样的方案，这里不在重复描述，为了适应数据集大小的变化，仅对 RNN 模型的 batch size 和 epoch 这两个参数进行了调整，batch size 参数调整为 512，epoch 调整为 100。图 7 为 PCA 和 t-SNE 可视化图，图 8 为生成数据与真实数据对比图，表 7 为模型训练数据。

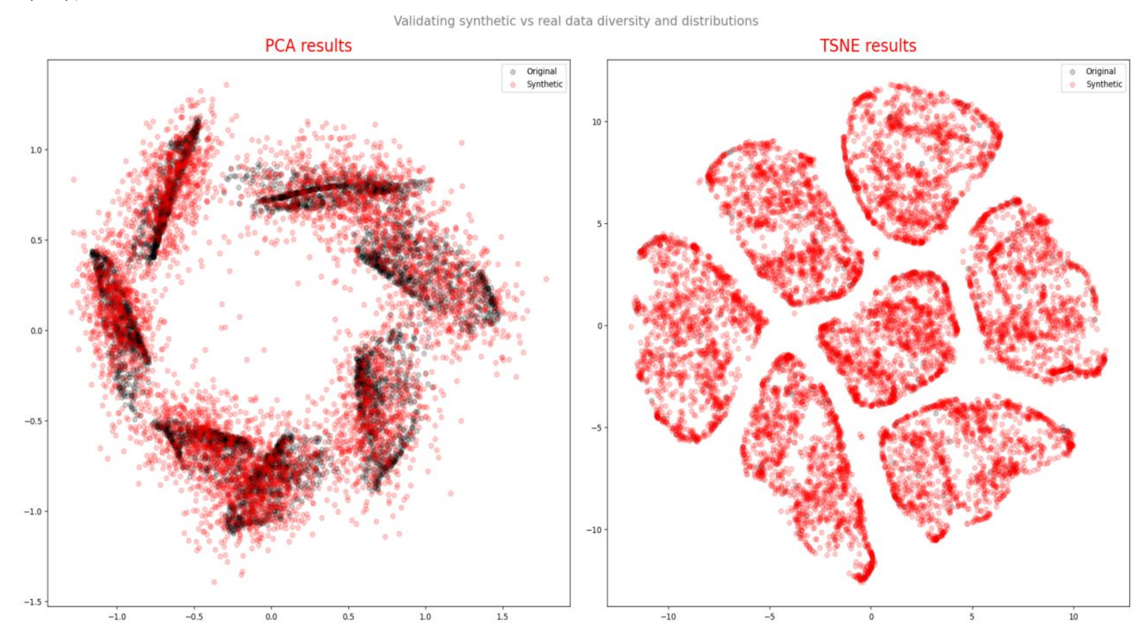


图 7 正样本 VS 正样本生成数据可视化对比图

表 7：正样本 VS 正样本生成数据模型表现

十轮均值	MSE	MAE	MSLE
真实数据	0.02674975	0.08859675	0.012944125
生成数据	0.030091625	0.1031815	0.01426
性能差距	-12.49%	-16.46%	-10.17%

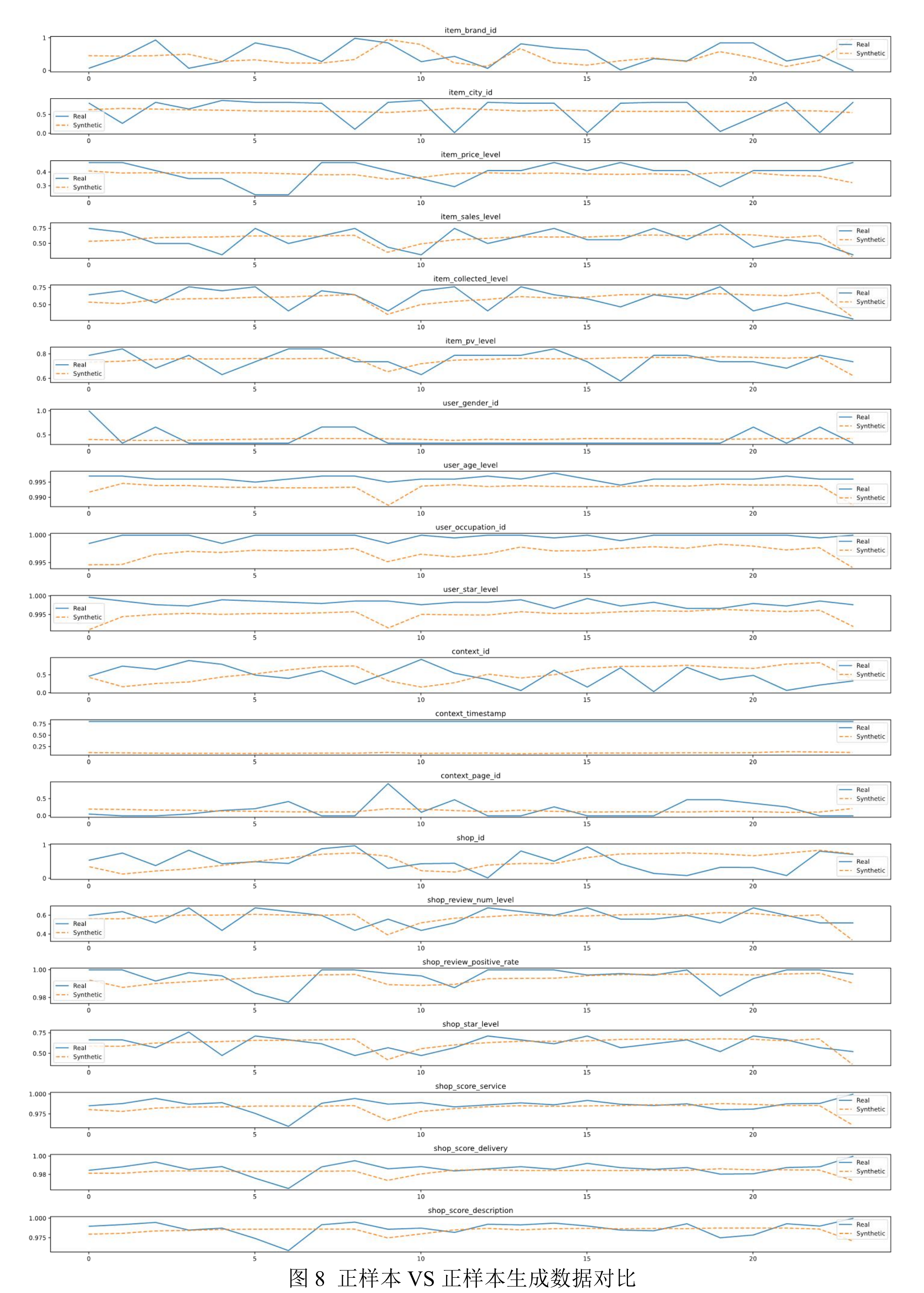


图 8 正样本 VS 正样本生成数据对比

通过以上的数据可视化和模型效果对比实验，可知 GAN 生成的数据和真实数据差距较小，可以用于模型实验。

2.5 实验结果

将上一步生成的数据进行整理后用于 XGBoost+LR 模型训练，将生成数据和经过标准化处理后的真实数据合并为同一个数据集（生成数据和真实数据大小一

致），进行训练和检验，和之前仅使用真实数据进行对比，具体数据如表 8 所示。

表 8：真实数据 VS 真实数据+生成数据模型表现

	真实数据	真实数据+生成数据	性能提升
logloss	0.072366	0.072177	0.26%
R2	0.069294	0.066981	3.34%
MSE	0.015079	0.015116	-0.25%
MAE	0.029819	0.030563	-2.50%
MSLE	0.007327	0.00725	1.05%

通过以上表格可以看出，使用生成数据可以改进模型的性能以及鲁棒性，但由于生成数据难以模仿生成真实数据中的极值，在一些指标上出现了性能损失，同时，可见此种方式更适合冷启动的初始模型训练，上线之后则要看业务实际需求使用。

3 总结与展望

本文在总结现有个性化推荐方法的基础上，结合工程实践中的成熟经验，提出利用基于 GAN+XGBoost+LR 的个性化推荐方法来解决长尾商品的冷启动，结合针对性的模型训练更新策略，在兼顾了性能和精确度要求的前提下提升了模型店性能和鲁棒性。下一步的研究方向是进一步深挖深度学习方法的使用，利用深度学习模型进一步替代 XGBoost 模型承担自动化特征工程部分功能，提高特征之间隐含关系的挖掘能力，优化 GAN 模型的设计，更符合数据特征的训练数据，进一步提高预测精度和鲁棒性。

参 考 文 献

[1] CAI Yifang,AII Jun,SU Zhan. Collaborative filtering personalized recommendation algorithm based on the quantity and quality of common scoring[J].Software Engineering,2022,25(08):20-24.

(蔡依芳,艾均,苏湛.融合共同评分数量和质量的协同过滤个性化推荐算法[J].软件工程,2022,25(08):20-24.)

[2] He X , Bowers S , Candela J Q , et al. Practical Lessons from Predicting Clicks on Ads at Facebook[C]// Eighth International Workshop on Data Mining for Online Advertising. ACM, 2014:1-9.

[3] LIU Taiheng. Research on personalized recommendation method based on deep learning[D].Guangdong University of Technology,2022.

(刘太亨.基于深度学习的个性化推荐方法研究[D].广东工业大学,2022.)

[4] ZAHID A, SHAREF N M, MUSTAPHA A. Normalization-based neighborhood model for cold start problem in recommendation system[J]. The International Arab Journal of Information Technology, 2020, 17(3): 281-290.

[5] TAHMASEBI F, MEGHDADI M, AHMADIAN S, et al. A hybrid recommendation system based on profile expansion technique to alleviate cold start problem[J]. Multi-media Tools and Applications, 2021， 80(2): 2339-2354

[6] PALETI L, FANG 丫 Approaching the cold-start problem using community detection based alternating least square factorization in recommendation systems[J]. Evolutionary Intelligence, 2021，14(2): 835 一 849.

[7] YADAV U, DUHAN N, BHATIA K K. Dealing with pure new user cold-start problem in recommendation system based on linked open data and social network features[J].Mobile Information Systems, 2020, 2020: 8912065 一 8912072.

[8] Alima,Tianchi Big Data Crowdwisdom Platform Alimama Search Ads Conversion Prediction [EB/OL].

https://tianchi.aliyun.com/competition/entrance/231647/introduction, 2018-3-1/2022-9-1.
(阿里妈妈，天池大数据众智平台．阿里妈妈搜索广告转化预测[EB/OL].https://tianchi.aliyun.com/competition/entrance/231647/introduction, 2018-3-1/2022-9-1.)

Yoon J,Jarrett D,Schaar M.Time-series Generative Adversarial Networks[C].Neural Information

通讯作者（Corresponding author）： 陈艳陈（Cheng Yanchun），
E-mail:376146485@qq.com

基金项目（）： 本文系“河北省推进数字经济与实体经济融合发展对策研究项目(21557623D)”基金项目的研究成果之一。
The work is supported by Hebei Province promotes the research project on countermeasures for the integrated development of digital economy and real economy (Grant No.21557623D).

作者贡献声明：

达钰鹏：提出了算法思路并进行实验，起草论文。
陈艳纯：论文最终版修订。

利益冲突声明：

所有作者声明不存在利益冲突关系。

支撑数据：

[1] 阿里妈妈，天池大数据众智平台．阿里妈妈搜索广告转化预测 .https://tianchi.aliyun.com/competition/entrance/231647/introduction, 2018-3-1/2022-9-1..